



CERTIFIED COPY OF  
PRIORITY DOCUMENT

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2000年 5月17日

出 願 番 号

Application Number:

特願2000-145168

出 願 人

Applicant(s):

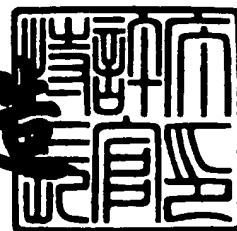
松下電器産業株式会社

CERTIFIED COPY OF  
PRIORITY DOCUMENT

2001年 5月18日

特許庁長官  
Commissioner,  
Japan Patent Office

及川耕造



【書類名】 特許願

【整理番号】 2033820044

【提出日】 平成12年 5月17日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 15/403  
G06F 15/40  
G06F 15/20

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式  
会社内

【氏名】 内藤 栄一

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式  
会社内

【氏名】 荒木 昭一

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式  
会社内

【氏名】 九津見 洋

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式  
会社内

【氏名】 小澤 順

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式  
会社内

【氏名】 丸野 進

【特許出願人】

【識別番号】 000005821

【氏名又は名称】 松下電器産業株式会社

【代理人】

【識別番号】 100097445

【弁理士】

【氏名又は名称】 岩橋 文雄

【選任した代理人】

【識別番号】 100103355

【弁理士】

【氏名又は名称】 坂口 智康

【選任した代理人】

【識別番号】 100109667

【弁理士】

【氏名又は名称】 内藤 浩樹

【手数料の表示】

【予納台帳番号】 011305

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9809938

【書類名】 明細書

【発明の名称】 情報検索装置、自動質問回答装置、情報検索方法および自動質問回答方法

【特許請求の範囲】

【請求項 1】 検索対象の複数の情報単位から検索条件に合致する情報単位を検索する情報検索装置であって、検索対象の情報単位を記憶する情報単位記憶手段と、前記情報単位の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前記情報単位のクラスタ分類を行うクラスタ分類手段と、前記情報単位から検索条件に合致する情報単位を検索する検索手段と、検索時に検索結果の情報単位が属する前記クラスタを出力する出力手段を有することを特徴とする情報検索装置。

【請求項 2】 前記情報単位は、文章の項目を少なくとも 1 つ含むことを特徴とする請求項 1 に記載の情報検索装置。

【請求項 3】 前記特徴量抽出手段は、前記情報単位の文章の項目から単語を切り出し、単語とその重みの組とからなる特徴量を抽出することを特徴とする請求項 2 に記載の情報検索装置。

【請求項 4】 前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位が属する前記クラスタの前記ラベルを出力することを特徴とする請求項 1 に記載の情報検索装置。

【請求項 5】 前記クラスタラベル作成手段は、前記特徴量に基づき前記クラスタの内容を表す単語からなるラベルを作成することを特徴とする請求項 4 に記載の情報検索装置。

【請求項 6】 前記クラスタラベル作成手段は、前記特徴量に基づき前記クラスタの内容を表す文からなるラベルを作成することを特徴とする請求項 4 に記載の情報検索装置。

【請求項 7】 前記特徴量に基づき前記情報単位の内容を表すラベルを作成する情報単位ラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位の前記ラベルと該情報単位が属するクラスタの前記ラベルとを出力することを特徴

とする請求項4に記載の情報検索装置。

【請求項8】 情報単位ラベル作成手段は、前記特徴量に基づき前記クラスタの内容を表す文からなるラベルを作成することを特徴とする請求項7に記載の情報検索装置。

【請求項9】 検索対象の複数の情報単位から検索条件に合致する情報単位を検索する情報検索方法であって、前記情報単位の特徴量を抽出し、前記特徴量に基づき前記情報単位のクラスタ分類を行い、前記クラスタの内容を表すラベルを作成し、前記特徴量に基づき前記情報単位の内容を表す文からなるラベルを作成し、前記情報単位から検索条件に合致する情報単位を検索し、検索結果の情報単位の前記ラベルと該情報単位が属するクラスタの前記ラベルとを出力することを特徴とする情報検索方法。

【請求項10】 複数の文書から利用者の質問に合致する文書を検索し利用者に回答する自動質問回答装置であって、回答対象の文書を記憶する文書記憶手段と、前記文書の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前記文書のクラスタ分類を行うクラスタ分類手段と、前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段と、前記特徴量に基づき前記文書の内容を表すラベルを作成する文書ラベル作成手段と、利用者からの質問と前記文書との類似度を算出する類似度算出手段と、前記類似度が所定の値以上の文書を回答文書として選択する回答選択手段と、前記回答文書の前記ラベルと該回答文書が属するクラスタの前記ラベルとを出力する出力手段とを有することを特徴とする自動質問回答装置。

【請求項11】 複数の文書から利用者の質問に合致する文書を検索し利用者に回答する自動質問回答方法であって、前記文書の特徴量を抽出し、前記特徴量に基づき前記文書のクラスタ分類を行い、前記特徴量に基づき前記クラスタの内容を表すクラスタラベルを作成し、前記特徴量に基づき前記文書の内容を表すラベルを作成し、利用者からの質問と前記文書との類似度を算出し、前記類似度が所定の値以上の文書を回答文書として選択し、前記回答文書の前記ラベルと該回答文書が属するクラスタの前記ラベルとを出力することを特徴とする自動質問回答方法。

## 【発明の詳細な説明】

## 【0001】

## 【発明の属する技術分野】

本発明は、大量の情報の中から利用者が求める情報を容易に見つけ出すことを可能とするための情報検索装置および、利用者からの質問に対して適切な回答を検索して回答する自動質問回答装置に関する。

## 【0002】

## 【従来の技術】

近年、インターネットの普及に伴い、WWW (World Wide Web) 上にHTML (Hyper Text Markup Language) で記述された様々なホームページが掲載されるようになるなど、一般利用者が大量の情報にアクセスすることが可能となっている。また、FAQ (Frequently Asked Questions) 集と称した、頻繁に問い合わせられる質問とその回答とを対にしたリストが公開されていて、利用者は質問に対する回答を得ることが可能である。これらの情報は、利用者にとって、求める情報の所在がわかれば、即座に閲覧できるので便利であるが、逆に大量の情報の中から自分の求める情報を見つけ出すことが大変な作業となっている。

## 【0003】

このため、文書からキーワードを切り出してその文書の特徴量とし、特徴量間の内積を算出して文書間の類似度を求め、質問文に対する類似文書を検索するという検索技術が知られている。

## 【0004】

## 【発明が解決しようとする課題】

しかしながら、インターネット上の情報や、あるいは、事例ベースで蓄積されたFAQ集は、多くの人が独立して情報を提供しているので、情報の重複は避けられず、同じような内容を持つ文書が大量に存在する。したがって、従来の技術では、質問文に類似した文書として、同じような内容の文書が大量に検索されてしまうことが多いので、利用者は結局大量の検索結果の中から欲しい情報を見つけ出す作業が必要であった。もしくは、検索結果を一定の数に制限すると、同じような内容の検索結果で占められてしまい、自分の欲しい情報がなかったりする

という課題があった。

【0005】

本発明は、かかる点に鑑み、同じような内容の検索結果をまとめて出力することにより、利用者に検索結果の内容を把握することを容易にし、情報検索の負担を軽減させる情報検索装置、自動質問回答装置、情報検索方法および自動質問回答方法を提供することを目的とする。

【0006】

【課題を解決するための手段】

請求項1の本発明は、検索対象の複数の情報単位から検索条件に合致する情報単位を検索する情報検索装置であって、検索対象の情報単位を記憶する情報単位記憶手段と、前記情報単位の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前記情報単位のクラスタ分類を行うクラスタ分類手段と、前記情報単位から検索条件に合致する情報単位を検索する検索手段と、検索時に検索結果の情報単位が属する前記クラスタを出力する出力手段を有することを特徴とする情報検索装置である。

【0007】

請求項4の本発明は、前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位が属する前記クラスタの前記ラベルを出力することを特徴とする請求項1に記載の情報検索装置である。

【0008】

請求項7の本発明は、前記特徴量に基づき前記情報単位の内容を表すラベルを作成する情報単位ラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位のラベルと該情報単位が属するクラスタの前記ラベルとを出力することを特徴とする請求項4に記載の情報検索装置である。

【0009】

請求項9の本発明は、複数の文書から利用者の質問に合致する文書を検索し利用者に回答する自動質問回答装置であって、回答対象の文書を記憶する文書記憶手段と、前記文書の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前

記文書のクラスタ分類を行うクラスタ分類手段と、前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段と、前記特徴量に基づき前記文書の内容を表すラベルを作成する文書ラベル作成手段と、利用者からの質問と前記文書との類似度を算出する類似度算出手段と、前記類似度が所定の値以上の文書を回答文書として選択する回答選択手段と、前記回答文書の前記ラベルと該回答文書が属するクラスタの前記ラベルとを出力する出力手段とを有することを特徴とする自動質問回答装置である。

【 0 0 1 0 】

【発明の実施の形態】

以下、本発明の実施の形態について図面を参照しながら説明する。

【 0 0 1 1 】

（実施の形態 1）

第 1 の実施の形態の機能構成図を図 1 に示す。図 1 において、101 は複数の文書を記憶する文書記憶部、102 は文書記憶部 101 に記憶されている文書から特徴ベクトルを抽出する特徴ベクトル抽出部、103 は特徴ベクトル抽出部 102 が求めた特徴ベクトルに基づき文書記憶部 101 に記憶されている文書のクラスタ分類を行うクラスタ分類部、104 はクラスタ分類部 103 がクラスタ分類した文書のクラスタを記憶するクラスタ記憶部、105 は文書記憶部 101 とクラスタ記憶部 104 とから文書を検索するデータベース検索部、106 は利用者との入出力を管理する利用者入出力部、107 は利用者からの入力を受け付ける入力部、108 は利用者に情報を表示する表示部である。

【 0 0 1 2 】

本実施の形態のハードウェア構成を図 2 に示す。図 2 は基本的に汎用の計算機システムの構成と同じである。また、図 1 に示した構成部分と同一の構成部分を含んでいるため、同一構成部分には同一番号を付して説明を省略する。図 2 において、112 はプログラムやデータを記憶する揮発性メモリからなる主記憶装置、113 はプログラムやデータを記憶する不揮発性メモリからなる補助記憶装置、111 は主記憶装置 112 に記憶されているプログラムを実行する CPU である。補助記憶装置 113 に記憶されているプログラムやデータは、主記憶装置 1

12にロードされた後CPU111により実行される。

【0013】

文書記憶部101には、検索の対象となる所与の $n$ 個の文書が記憶されている( $n \geq 2$ )。記憶されている文書の例を図3に示す。文書は、ユニークな文書ID、文章形式の本文とからなる。 $i$ 番目の文書を $D_i$ とする( $1 \leq i \leq n$ )。

【0014】

以上のように構成された第1の実施の形態の動作を、文書登録時と文書検索時に分けて以下に説明する。文書登録時とは、初めて文書が文書記憶部101に登録される場合、あるいは、それ以降に文書の追加／変更／削除があった場合に行われる動作である。文書検索時とは、登録されている文書を検索して閲覧する場合に行われる動作である。

【0015】

「文書登録時」

まず、特徴ベクトル抽出部102は、文書記憶部101に記憶されている全ての文書 $D_i$ を取り出し、各文書の特徴ベクトル $V_i$ を抽出する。特徴ベクトルは、文書の特徴を表す単語 $t_j$ とその重み $w_{ij}$ の組を要素とするベクトルであり、その要素の数は文書によって異なる。ここで、 $j$ は単語を識別するユニークな番号である。抽出された特徴ベクトル $V_i$ の例を図4に示す。特徴ベクトルの抽出手順を図5に示すフローチャートを用いて説明する。

【0016】

[ステップ101] カウンタの初期化

文書のカウンタ $i$ に $i=1$ を設定する。

【0017】

[ステップ102] 単語の抽出

文書記憶部101から文書 $D_i$ を取り出し、形態素解析、構文解析、不要語除去など、一般に知られている方法により、本文から出現する単語 $t_j$ を抽出し、 $t_j$ の文書内での出現回数 $f_{ij}$ をカウントする。

【0018】

[ステップ103] 終了判定

全文書につき[ステップ102]の処理が終了した場合、すなわち*i*=*n*の場合[ステップ105]に進む。そうでない場合[ステップ104]に進む。

【0019】

[ステップ104] カウンタの増加

カウンタ*i*を1増加し[ステップ102]に進む。

【0020】

[ステップ105] idf値算出

単語 $t_j$ の全文書に対する重要度として、単語 $t_j$ が出現する文書数の少なさを表す度合idf(inverse document frequency)値を(数1)により算出する。

【0021】

【数1】

$$idf_j = \log \frac{n}{\text{語}t_j\text{が出現する文書数}} + 1$$

【0022】

[ステップ106] カウンタの初期化

文書のカウンタ*i*に*i*=1を設定する。

【0023】

[ステップ107] tf・idf値算出

単語 $t_j$ が文書 $D_i$ を特徴付ける重み $w_{ij}$ として、単語 $t_j$ の文書 $D_i$ 内での出現割合 $t$ f(term frequency)値とidf値とを掛け合わせたtf・idf値を(数2)により算出する。

【0024】

【数2】

$$w_{ij} = \frac{f_{ij}}{\sum_{j:t_j \in D_i} f_{ij}} \cdot idf_j$$

【0025】

[ステップ108] 終了判定

全文書につき[ステップ107]の処理が終了した場合、すなわち $i=n$ の場合終了する。そうでない場合[ステップ109]に進む。

【0026】

[ステップ109] カウンタの増加

カウンタ $i$ を1増加し[ステップ102]に進む。

【0027】

次に、クラスタ分類部103は、特徴ベクトル抽出部102が抽出した特徴ベクトルを用いて、全ての文書を $m$ 個のクラスタに分類する( $1 < m < n$ )。ここで、 $k$ 番目のクラスタを $C_k$ とする( $1 \leq k \leq m$ )。クラスタ分類の手順として、樹形図的に逐次クラスタに分類していく階層的クラスタリングを用い、図6を用いて説明する。

【0028】

[ステップ111] クラスタ間初期距離計算

クラスタの初期値として、1つの文書 $D_i$ だけを要素として持つクラスタ $C_i = [D_i]$ を設定する。すなわち初期クラスタは $n$ 個ある。各クラスタ $C_k, C_l$  ( $1 \leq k, l \leq n$ )間の距離 $d_{kl}$ として、各文書の特徴ベクトル間の距離を類似比を用いて(数3)により算出する。(数3)において、 $\wedge$ はmin演算、 $\vee$ はmax演算を表す。

【0029】

【数3】

$$d_{kl} = -\log \frac{\sum_{j: x_j \in D_k \cup D_l} w_{kj} \wedge w_{lj}}{\sum_{j: x_j \in D_k \cup D_l} w_{kj} \vee w_{lj}}$$

【0030】

[ステップ112] カウンタの初期化

クラスタリング回数のカウンタ $i$ に $i=1$ を設定する。

## 【0031】

[ステップ113] 距離最小クラスタ探索

全てのクラスタの組み合わせの中で、クラスタ間距離 $D_{kl}$ が最も小さいクラスタ $C_k, C_l (k < l)$ の組を探索する。

## 【0032】

[ステップ114] クラスタ統合

クラスタ $C_k, C_l$ を統合してクラスタ $C_g$ とする。すなわち、 $C_g = C_k \cup C_l$ 、 $C_l = \phi$ とする( $\phi$ は空集合を表す)。クラスタの統合に伴い、クラスタ $C_g$ と他のクラスタ $C_h (1 \leq h \leq n)$ とのクラスタ間距離をワード法を用いて(数4)により算出する。

(数4)において $n_k$ はクラスタ $C_k$ の要素の数である。

## 【0033】

【数4】

$$d_{gh} = \frac{(n_k + n_h) \cdot d_{kh} + (n_l + n_h) \cdot d_{lh} - n_h \cdot d_{kl}}{n_g + n_h}$$

## 【0034】

[ステップ115] 終了判定

クラスタリング回数が $n-1$ の場合、すなわち全ての初期クラスタが1つのクラスタに統合された場合[ステップ117]に進む。そうでない場合[ステップ116]に進む。

## 【0035】

[ステップ116] カウンタの増加

カウンタ $i$ を1増加し[ステップ112]に進む。

## 【0036】

[ステップ117] クラスタ数決定

[ステップ111]から[ステップ115]までのクラスタ分類過程においては、クラスタリング回数ごとにクラスタの数は1つずつ減少する。このステップでは、クラスタ分類過程を振り返り、適切なクラスタリング回数を決定する。ここでは、要素を2つ以上持つクラスタの数が最大になるクラスタリング回数を適切なクラス

タリング回数であるとする。

【0037】

[ステップ118] クラスタ要素書き出し

[ステップ117]で決定したクラスタリング回数までクラスタ分類を行った時点での各クラスタに含まれる要素をクラスタ記憶部104に書き出す。クラスタ記憶部104に書き出されたクラスタの例を図7に示す。各クラスタは、クラスタIDとそのクラスタに含まれる文書の文書IDとからなる。例えば、クラスタ1には、1, 190, 432, 644番の4つの文書が含まれている。これは、この4つの文書の特徴ベクトル同士が、他の文書にくらべて類似していることを表している。

【0038】

以上の動作により、文書登録時に、文書のクラスタを作成して記憶しておく。

【0039】

「文書検索時」

まず、利用者入出力部106は、入力部107を通じて文書の検索条件を受け付ける。検索条件としては、文書のキーワード、文書IDなど文書検索の条件となるものなら何でもよい。

【0040】

データベース検索部105は、文書記憶部101から検索条件を満たす文書を検索し、次に、クラスタ記憶部104から検索された文書が含まれるクラスタを検索し、最後に、再び文書記憶部101から検索されたクラスタに含まれる文書を検索し、その結果を利用者入出力部106に送る。

【0041】

利用者入出力部106は、表示部108を通じて検索結果を利用者に表示する。検索結果の例を図8に示す。図8では、クラスタIDと、そのクラスタに含まれる文書の文書IDと本文とを、クラスタごとに表形式で表示し、マウスで「前のクラスタ」ボタンや「次のクラスタ」ボタンを押して別のクラスタを表示することにより、全ての検索結果を表示することができる。

【0042】

以上のように、第1の実施の形態によれば、文書の特徴ベクトルを抽出し、特

徴ベクトルに基づいて文書をクラスタ分類して記憶し、文書の検索結果をクラスタごとにまとめて表示することにより、検索結果を類似した文書の固まりとして把握することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

## 【 0 0 4 3 】

なお、本実施の形態では、文書は所与のものがあらかじめ記憶されていたが、光ディスクなどの記憶媒体やインターネットなどのネットワーク媒体などにより、後から新たに導入、もしくは、改訂されても良い。また、対象情報特徴ベクトルの算出は、 $tf \cdot idf$  値を用いていたが、単純に単語の出現回数とするなど、他の方法でも良い。また、クラスタ分類の方法として、階層的クラスタリングを用いたが、非階層的クラスタリングでも良い。また、初期クラスタ間距離として（数 3）の類似比を用いたが、ユークリッド平方距離など他の距離を用いても良い。また、クラスタ統合時のクラスタ間距離の算出手法として（数 4）のワード法を用いたが、最長距離法など他の手法を用いても良い。また、クラスタ数の決定手法として、要素を 2 つ以上持つクラスタの数が最大になるクラスタリング回数としたが、クラスタ数を文書数の一定の割合とするなど他の決定手法でも良い。また、文書の検索は、キーワードや文書 ID によるもの以外に、全文検索であってもあいまい検索であっても良い。また、検索結果として文書 ID をも表示したが、表示しなくても良い。

## 【 0 0 4 4 】

## （実施の形態 2）

次に本発明の第 2 の実施の形態について説明する。第 1 の実施の形態では、検索結果の文書をクラスタにまとめて表示する例で説明したが、第 2 の実施の形態は、クラスタに単語からなる単語ラベルを付けて、クラスタの内容を把握し易くできるように考慮したものである。

## 【 0 0 4 5 】

第 2 の実施の形態の機能構成図を図 9 に示す。図 9 において、図 1 に示した第 1 の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第 2 の実施の形態と第 1 の実施の形態との相違点は、クラスタ分類部 1 0 3 が作成したク

ラスタについて単語からなる単語ラベルを作成するクラスタ単語ラベル作成部 201 と、クラスタ単語ラベル作成部 201 が作成した単語ラベルを記憶するクラスタラベル記憶部 202 とを追加したことである。

【0046】

文書記憶部 101 に記憶されている文書の例として、第 1 の実施の形態と同じ図 3 に示したものをを用いる。

【0047】

以上のように構成された第 2 の実施の形態の動作を、第 1 の実施の形態と同様に、文書登録時と文書検索時に分けて以下に説明する。

【0048】

「文書登録時」

クラスタ分類部 103 がクラスタ分類を行うまでの処理は、第 1 の実施の形態の動作と同様であるので説明は省略する。それ以降のクラスタ単語ラベル作成部 201 の動作を図 10 のフローチャートを用いて説明する。

【0049】

[ステップ201] カウンタの初期化

クラスタのカウンタ  $k$  に  $k=1$  を設定する。

【0050】

[ステップ202] 単語出現回数計数

クラスタ  $C_k$  の要素である全ての文書  $D_i$  に含まれる単語  $t_j$  ごとに、単語  $t_j$  が含まれる文書の数のカウントする。すなわち、単語  $t_j$  のクラスタ  $C_k$  内での共起回数を数える。

【0051】

[ステップ203]  $tf \cdot idf$  値積算

クラスタ  $C_k$  の要素である全ての文書  $D_i$  に含まれる単語  $t_j$  ごとに、単語  $t_j$  の  $tf \cdot idf$  値  $w_{ij}$  の合計を算出する。

【0052】

[ステップ204] 単語のソート

クラスタ  $C_k$  の要素である全ての文書  $D_i$  に含まれる全ての単語  $t_j$  を、[ステップ2

02]で求めた共起回数の多い順にソートする。回数が同じ場合は[ステップ202]で求めたtf・idf値の合計の大きい順にソートする。

#### 【 0 0 5 3 】

[ステップ205] クラスタ単語ラベルの書き出し

[ステップ204]でソートされた上位の3つの単語を選択し、クラスタの単語ラベルとしてクラスタ記憶部202に書き出す。クラスタラベル記憶部に書き出された単語ラベルの例を図11に示す。例えば、クラスタ1には、「お菓子」「間食」「チーズ」という単語ラベルが付いていることを表す。

#### 【 0 0 5 4 】

[ステップ206] 終了判定

全クラスタにつき[ステップ202]から[ステップ205]までの処理が終了した場合、すなわち $k=m$ の場合終了する。そうでない場合[ステップ207]に進む。

#### 【 0 0 5 5 】

[ステップ207] カウンタの増加

カウンタ $k$ を1増加し[ステップ202]に進む。

#### 【 0 0 5 6 】

以上の動作により、文書登録時に、文書のクラスタ、クラスタの単語ラベルとを作成して記憶しておく。

#### 【 0 0 5 7 】

「文書検索時」

データベース検索部105がクラスタに含まれる文書を検索するまでの処理は、第1の実施の形態の動作と同様であるので説明は省略する。クラスタに含まれる文書を検索した後、データベース検索部105は、クラスタラベル記憶部202から、検索されたクラスタの単語ラベルを検索し、その結果を利用者入出力部106と表示部108を通じて利用者に表示する。検索結果の例を図12に示す。図12と第1の実施の形態での表示例の図8との違いは、クラスタにクラスタ単語ラベルが表示されていることである。

#### 【 0 0 5 8 】

以上のように、第2の実施の形態によれば、文書の特徴ベクトルを抽出し、特

徴ベクトルに基づいて文書をクラスタ分類し、単語からなるクラスタの単語ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて単語ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を理解することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

## 【0059】

なお、本実施の形態では、単語ラベルの作成方法として単語の共起回数でソートしたが、 $tf \cdot idf$ 値のみでソートするなど他の方法でも良い。また、単語ラベルの単語数を3つにしたが、3つ以外でも良い。また、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

## 【0060】

## (実施の形態3)

次に本発明の第3の実施の形態について説明する。第2の実施の形態では、単語からなるクラスタの単語ラベルを表示する例で説明したが、第3の実施の形態は、クラスタに文からなる文ラベルを付けて、クラスタの内容を自然文で把握し易くできるように考慮したものである。

## 【0061】

第3の実施の形態の機能構成図を図13に示す。図13において、図9に示した第2の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第3の実施の形態と第2の実施の形態との相違点は、クラスタ単語ラベル作成部201を、クラスタ分類部103が作成したクラスタについて文からなる文ラベルを作成するクラスタ文ラベル作成部301に置き換えたことである。

## 【0062】

文書記憶部101に記憶されている文書の例として、第1の実施の形態と同じ図3に示したものをを用いる。

## 【0063】

以上のように構成された第3の実施の形態の動作を、第1の実施の形態と同様に、文書登録時と文書検索時に分けて以下に説明する。

## 【0064】

## 「文書登録時」

クラスタ分類部 1 0 3 がクラスタ分類を行うまでの処理は、第 1 の実施の形態の動作と同様であるので説明は省略する。それ以降のクラスタ文ラベル作成部 3 0 1 の動作を図 1 4 のフローチャートを用いて説明する。

## 【 0 0 6 5 】

[ステップ301] カウンタの初期化

クラスタのカウンタ  $k$  に  $k=1$  を設定する。

## 【 0 0 6 6 】

[ステップ302] 単語出現回数計数

第 2 の実施の形態と同様に、クラスタ  $C_k$  の要素である全ての文書  $D_i$  に含まれる単語  $t_j$  ごとに、単語  $t_j$  が含まれるクラスタ内の文書の数をカウントする。すなわち、単語  $t_j$  のクラスタ  $C_k$  内での共起回数を数える。

## 【 0 0 6 7 】

[ステップ303] 文ごとに積算

クラスタ  $C_k$  の要素である全ての文書  $D_i$  を構成する文ごとに、[ステップ302] でカウントした単語の共起回数の合計を算出する。ここで、文とは、文書を「。」などの句点で区切った1つ1つをいう。

## 【 0 0 6 8 】

[ステップ304] 文のソート

クラスタ  $C_k$  の要素である全ての文書  $D_i$  を構成する文を [ステップ303] で求めた共起回数の合計の大きい順にソートする。

## 【 0 0 6 9 】

[ステップ305] クラスタ文ラベルの書き出し

[ステップ304] でソートされた最上位の文を選択し、クラスタの文ラベルとしてクラスタ記憶部 2 0 2 に書き出す。クラスタラベル記憶部 2 0 2 に書き出された文ラベルの例を図 1 5 に示す。例えば、クラスタ1には、「水分の多い物（ゼリー、プリン、ヨーグルト）を…」という文ラベルが付いていることを表す。

## 【 0 0 7 0 】

[ステップ306] 終了判定

全クラスタにつき[ステップ302]から[ステップ305]までの処理が終了した場合、すなわち $k=m$ の場合終了する。そうでない場合[ステップ307]に進む。

【0071】

[ステップ307] カウンタの増加

カウンタ $k$ を1増加し[ステップ202]に進む。

【0072】

以上の動作により、文書登録時に、文書のクラスタ、クラスタの文ラベルとを作成して記憶しておく。

【0073】

「文書検索時」

データベース検索部105が検索を行って検索結果を利用者に表示するまでの動作は、第2の実施の形態と同様である。検索結果の例を図16に示す。図16と第2の実施の形態での表示例の図12との違いは、クラスタのラベルが単語ではなく文で表現されていることである。

【0074】

以上のように、第3の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、文からなるクラスタの文ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて文ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を自然文として理解することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

【0075】

なお、本実施の形態では、文ラベルの作成方法として単語の共起回数を合計したが、 $tf \cdot idf$ 値を合計するなど他の方法でも良い。また、単語ラベルの単語数を3つにしたが、3つ以外でも良い。また、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【0076】

(実施の形態4)

次に本発明の第4の実施の形態について説明する。第3の実施の形態では、ク

ラストの文ラベルを表示する例で説明したが、第4の実施の形態は、クラスタの文ラベルだけではなく、クラスタの要素である文書にも文による文書ラベルを付けて、文書の内容をも自然文で把握し易くできるように考慮したものである。

#### 【0077】

第4の実施の形態の機能構成図を図17に示す。図17において、図13に示した第3の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第4の実施の形態と第3の実施の形態との相違点は、クラスタ分類部103が作成したクラスタの要素である文書について文からなる文書ラベルを作成する文書ラベル作成部401と、文書ラベル作成部401が作成した文書ラベルを記憶する文書ラベル記憶部402を追加したことである。

#### 【0078】

文書記憶部101に記憶されている文書の例として、第1の実施の形態と同じ図3に示したものをを用いる。

#### 【0079】

以上のように構成された第4の実施の形態の動作を、第1の実施の形態と同様に、文書登録時と文書検索時に分けて以下に説明する。

#### 【0080】

##### 「文書登録時」

クラスタ文ラベル作成部301がクラスタの文ラベルを作成してクラスタラベル記憶部に書き出すまでは、第3の実施の形態の動作と同様であるので説明は省略する。それ以降の文書ラベル作成部401の動作を図18のフローチャートを用いて説明する。

#### 【0081】

[ステップ401] カウンタの初期化

文書のカウンタ $i$ に $i=1$ を設定する。

#### 【0082】

[ステップ402]  $tf \cdot idf$ 値積算

文書 $D_i$ を構成する各文ごとに、その文に含まれる全単語 $t_j$ の $tf \cdot idf$ 値 $w_{ij}$ の合計を算出する。

【0083】

[ステップ403] 終了判定

全文書につき[ステップ402]の処理が終了した場合、すなわち $i=n$ の場合終了する。そうでない場合[ステップ404]に進む。

【0084】

[ステップ404] カウンタの増加

カウンタ $i$ を1増加し[ステップ402]に進む。

【0085】

[ステップ405] カウンタの初期化

クラスタのカウンタ $k$ に $k=1$ を設定する。

【0086】

[ステップ406] 文のソート

クラスタ $C_k$ の要素である全ての文書 $D_i$ を構成する文を、[ステップ402]で求めた合計の多い順にソートする。

【0087】

[ステップ407] 文書ラベル選択

文書 $D_i$ の文書ラベルとして[ステップ406]でソートされた最上位の文を選択する。ただし、選択された文が、クラスタ文ラベル作成部が作成したクラスタの文ラベルと同一の場合は、文書 $D_i$ の文書ラベルとして[ステップ406]でソートされた上位から2番目の文を選択する。

【0088】

[ステップ408] 文書ラベルの書き出し

[ステップ406]で選択された文書 $D_i$ の文ラベルを文書ラベル記憶部402に書き出す。文書ラベル記憶部402に書き出された文書ラベルの例を図19に示す。例えば、クラスタ1に含まれる文書1には、「かみごたえがあり、後を引かないもので、…」という文書ラベルが付いていることを表す。

【0089】

[ステップ409] 終了判定

全クラスタにつき[ステップ406]から[ステップ408]までの処理が終了した場

合、すなわち $k=m$ の場合終了する。そうでない場合[ステップ410]に進む。

【0090】

[ステップ410] カウンタの増加

カウンタ $k$ を1増加し[ステップ406]に進む。

【0091】

「文書検索時」

データベース検索部105がクラスタ文ラベルを検索するまでの動作は、第3の実施の形態と同様であるので説明は省略する。クラスタ文ラベルの検索後、データベース検索部105は、文書ラベル記憶部402から、検索された文書の文書ラベルを検索し、その結果を利用者入出力部106と表示部108を通じて利用者に表示する。検索結果の例を図20に示す。図20と第3の実施の形態での表示例の図16との違いは、文書に文書ラベルである文が下線付きで表示されていることである。

【0092】

以上のように、第4の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各文書の内容を理解することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

【0093】

なお、本実施の形態では、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【0094】

(実施の形態5)

次に本発明の第5の実施の形態について説明する。第5の実施の形態は、自由文による質問から、過去の事例から検索して適切な回答を返答する自動質問回答システムの例である。第5の実施の形態の機能構成図を図21に示す。図21に

において、図17に示した第4の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第5の実施の形態と第4の実施の形態との相違点は、特徴ベクトル抽出部501が抽出した特徴ベクトルを記憶する特徴ベクトル記憶部501と、利用者の検索質問文と特徴ベクトル記憶部501が記憶している文書の特徴ベクトルとの類似度を求める類似度演算部502とを追加したことである。

## 【0095】

文書記憶部101には、検索の対象となる所与の複数の文書が記憶されている。記憶されている文書の例を図22に示す。文書は、ユニークな文書ID、文章形式の質問と質問に対する回答との事例からなる。

## 【0096】

以上のように構成された第5の実施の形態の動作を、第1の実施の形態と同様に、文書登録時と質問回答時に分けて以下に説明する。

## 【0097】

## 「文書登録時」

まず、特徴ベクトル抽出部102は、文書記憶部101に記憶されている全ての文書 $D_i$ から、各文書の質問と回答のそれぞれの特徴ベクトル $V_{Qi}$ 、 $V_{Ai}$ を抽出し、抽出された特徴ベクトルを特徴ベクトル記憶部501に書き出す。特徴ベクトルの抽出手順は第1の実施の形態と同様であるので説明を省略する。第1の実施の形態との違いは、文書の質問と回答の部分についてそれぞれ特徴ベクトルを算出する点と、特徴ベクトルを特徴ベクトル記憶部501に書き出す点である。

## 【0098】

次に、クラスタ分類部103は、特徴ベクトル記憶部501から回答の特徴ベクトル $V_{Ai}$ を読み出し、全ての文書をクラスタ $C_k$ に分類し、クラスタ記憶部104にクラスタを書き出す。クラスタ分類の手順は第1の実施の形態と同様であるので説明を省略する。第1の実施の形態との違いは、回答の特徴ベクトルを用いてクラスタ分類を行う点である。分類されたクラスタの例として第1の実施の形態と同じ図7のものをを用いる。

## 【0099】

クラスタ文ラベル作成部301は、各クラスタの文ラベルを作成し、クラスタ

ラベル記憶部202に書き出す。クラスタの文ラベルの作成手順は第3の実施の形態と同様であるので説明を省略する。作成された文ラベルの例として第3の実施の形態と同じ図15のものをを用いる。

#### 【0100】

文書ラベル作成部401は、各文書の文書ラベルを作成し、文書ラベル記憶部402に書き出す。文書ラベルの作成手順は第4の実施の形態と同様であるので説明を省略する。作成された文書ラベルの例として第4の実施の形態と同じ図19のものをを用いる。

#### 【0101】

以上の動作により、文書登録時に、質問と回答についてそれぞれ特徴ベクトルを作成し、また、回答について、クラスタ、クラスタ文ラベル、文書ラベルを作成し、それぞれの記憶部に記憶しておく。

#### 【0102】

##### 「文書検索時」

まず、利用者入出力部106は、入力部107を通じて、自由文による利用者からの質問Qを受け付ける。

#### 【0103】

特徴ベクトル抽出部102は、質問Qの特徴ベクトル $V_Q$ を抽出する。特徴ベクトルの抽出手順を図23に示すフローチャートを用いて説明する。

#### 【0104】

##### [ステップ501] 単語の抽出

質問Qから出現する単語 $t_j$ を抽出し、 $t_j$ の文書内での出現回数 $f_{ij}$ をカウントする。単語を抽出する方法は、第1の実施の形態と同様である。

#### 【0105】

##### [ステップ502] idf値算出

$t_j$ のidf値を算出する。idf値として、単語 $t_j$ が文書記憶部101のいずれかの文書中に存在する場合は、 $idf_j$ が既に文書登録時に算出されているので、それを用いる。単語 $t_j$ が存在しない場合は（数5）により算出する。

#### 【0106】

【数5】

$$idf_j = \log(n+1) + 1$$

【0107】

[ステップ503] tf・idf値算出

単語 $t_j$ のtf・idf値を算出する。tf・idf値算出の方法は、第1の実施の形態と同様である。抽出された質問の特徴ベクトルの例を図24に示す。

【0108】

類似度演算部502は、特徴ベクトル記憶部501から全ての文書の質問の特徴ベクトル $V_{Qi}$ を取り出し、文書の質問の特徴ベクトル $V_{Qi}$ と利用者からの質問の特徴ベクトル $V_Q$ との類似度を算出する。類似度の算出手順を図25に示すフローチャートを用いて説明する。

【0109】

[ステップ511] カウンタの初期化

文書のカウンタ $i$ に $i=1$ を設定する。

【0110】

[ステップ512] 特徴ベクトル内積演算

特徴ベクトル $V_{Qi}$ と利用者からの質問の特徴ベクトル $V_Q$ との類似度 $e_i$ を（数6）によりベクトルの内積で算出する。

【0111】

【数6】

$$e_i = V_{Qi} \cdot V_Q = \frac{\sum_j w_{ij} \cdot w_{Qj}}{|V_{Qi}| \cdot |V_Q|}$$

【0112】

[ステップ513] 終了判定

全文書につき[ステップ512]の処理が終了した場合、すなわち $i=n$ の場合[ステップ515]に進む。そうでない場合[ステップ514]に進む。

【0 1 1 3】

[ステップ514] カウンタの増加

カウンタ $i$ を1増加し[ステップ512]に進む。

【0 1 1 4】

[ステップ515] 文書のソート

全ての文書 $D_i$ を、[ステップ512]で求めた類似度 $e_i$ の高い順にソートする。

【0 1 1 5】

データベース検索部105は、文書記憶部101から、類似度演算部502が算出した類似度 $e_i$ が最上位の文書 $D_i$ を検索し、次に、クラスタ記憶部104から文書 $D_i$ が含まれるクラスタ $C_k$ を検索し、最後に、再び文書記憶部101からクラスタ $C_k$ に含まれる文書を検索し、その結果を利用者入出力部106に送る。

【0 1 1 6】

利用者入出力部106は、表示部108を通じて検索結果を利用者に表示する。検索結果の例として、第4の実施の形態と同じ図20のものを用いる。

【0 1 1 7】

以上のように、第5の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、利用者からの質問が入力された場合に、特徴ベクトルに基づいて類似文書を検索し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、質問に対する回答を類似した文書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各回答の内容を理解することが容易な自動質問回答システムを提供することが可能となる。

【0 1 1 8】

なお、本実施の形態では、特徴ベクトルの類似度演算方法としてベクトルの内積を用いたが、ベクトルの類似比を用いるなど他の方法でも良い。また、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【0 1 1 9】

【発明の効果】

第 1 の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類して記憶し、文書の検索結果をクラスタごとにまとめて表示することにより、検索結果を類似した文書の固まりとして把握することが容易となる。

## 【 0 1 2 0 】

第 2 の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、単語からなるクラスタの単語ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて単語ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を理解することが容易となる。

## 【 0 1 2 1 】

第 3 の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、文からなるクラスタの文ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて文ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を自然文として理解することが容易となる。

## 【 0 1 2 2 】

第 4 の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各文書の内容を理解することが容易となる。

## 【 0 1 2 3 】

第 5 の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、利用者からの質問が入力された場合に、特徴ベクトルに基づいて類似文書を検索し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、質問に対する回答を類似した文

書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各回答の内容を理解することが容易な自動質問回答システムを提供することが可能となる。

【図面の簡単な説明】

【図 1】

本発明第 1 の実施の形態の機能構成図

【図 2】

本発明第 1 の実施の形態のハードウェア構成を示す図

【図 3】

本発明第 1 の実施の形態の文書の例を示す図

【図 4】

本発明第 1 の実施の形態の文書の特徴ベクトルの例を示す図

【図 5】

本発明第 1 の実施の形態の特徴ベクトルの抽出手順を示すフローチャート

【図 6】

本発明第 1 の実施の形態のクラスタ分類手順を示すフローチャート

【図 7】

本発明第 1 の実施の形態のクラスタの例を示す図

【図 8】

本発明第 1 の実施の形態の検索結果の例を示す図

【図 9】

本発明第 2 の実施の形態の機能構成図

【図 1 0】

本発明第 2 の実施の形態のクラスタ単語ラベルの作成手順を示すフローチャート

【図 1 1】

本発明第 2 の実施の形態のクラスタ単語ラベルの例を示す図

【図 1 2】

本発明第 2 の実施の形態の検索結果の例を示す図

【図 1 3】

本発明第 3 の実施の形態の機能構成図

【図 1 4】

本発明第 3 の実施の形態のクラスタ文ラベルの作成手順を示すフローチャート

【図 1 5】

本発明第 3 の実施の形態のクラスタ文ラベルの例を示す図

【図 1 6】

本発明第 3 の実施の形態の検索結果の例を示す図

【図 1 7】

本発明第 4 の実施の形態の機能構成図

【図 1 8】

本発明第 4 の実施の形態の文書ラベルの作成手順を示すフローチャート

【図 1 9】

本発明第 4 の実施の形態の文書ラベルの例を示す図

【図 2 0】

本発明第 4 の実施の形態の検索結果の例を示す図

【図 2 1】

本発明第 5 の実施の形態の機能構成図

【図 2 2】

本発明第 5 の実施の形態の文書の例を示す図

【図 2 3】

本発明第 5 の実施の形態の利用者の質問の特徴ベクトルの抽出手順を示すフローチャート

【図 2 4】

本発明第 5 の実施の形態の利用者の質問の特徴ベクトルの例を示す図

【図 2 5】

本発明第 5 の実施の形態の類似度演算の手順を示すフローチャート

【符号の説明】

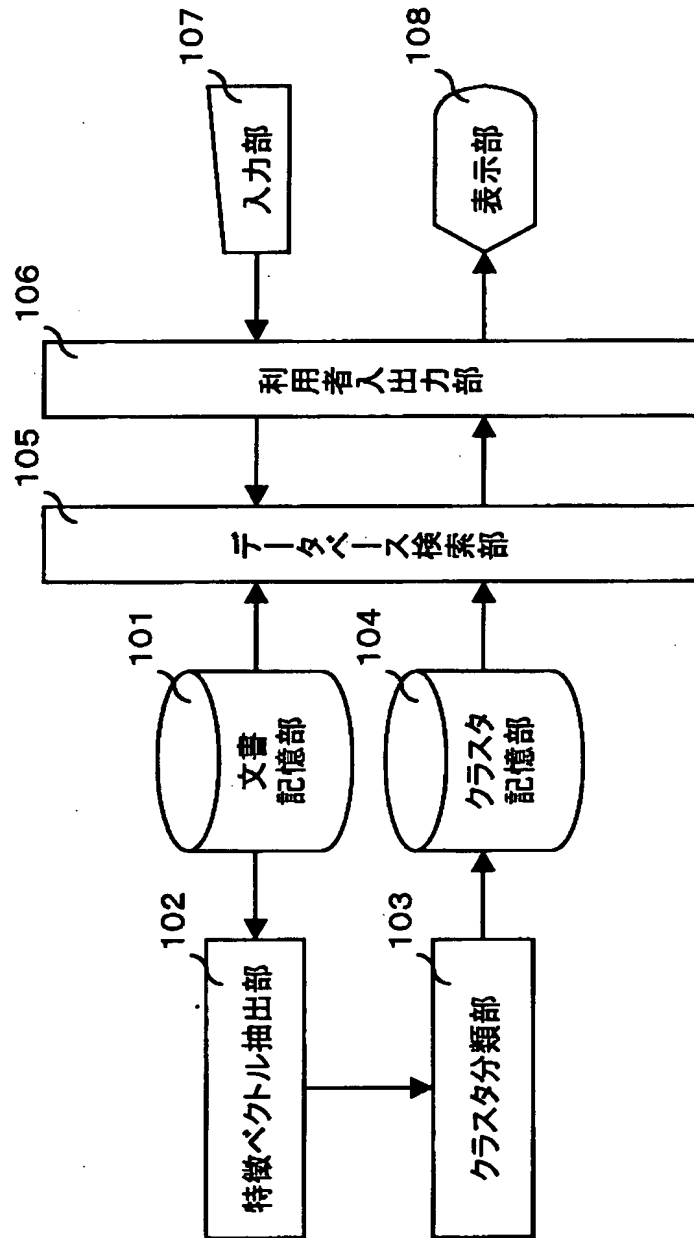
1 0 1 文書記憶部

- 1 0 2 特徴ベクトル抽出部
- 1 0 3 クラスタ分類部
- 1 0 4 クラスタ記憶部
- 1 0 5 データベース検索部
- 1 0 6 利用者入出力部
- 1 0 7 入力部
- 1 0 8 表示部
- 1 1 1 C P U
- 1 1 2 主記憶装置
- 1 1 3 補助記憶装置
- 2 0 1 クラスタ単語ラベル作成部
- 2 0 2 クラスタラベル記憶部
- 3 0 1 クラスタ文ラベル作成部
- 4 0 1 文書ラベル作成部
- 4 0 2 文書ラベル記憶部
- 5 0 1 特徴ベクトル記憶部
- 5 0 2 類似度演算部

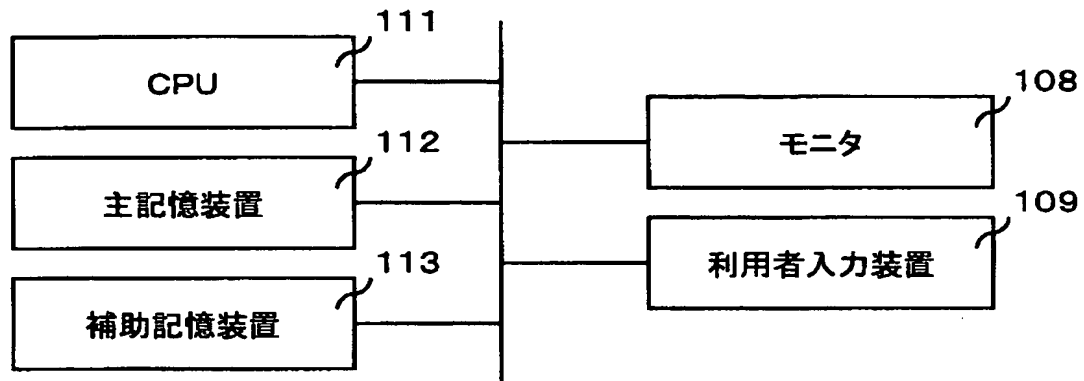
【書類名】

図面

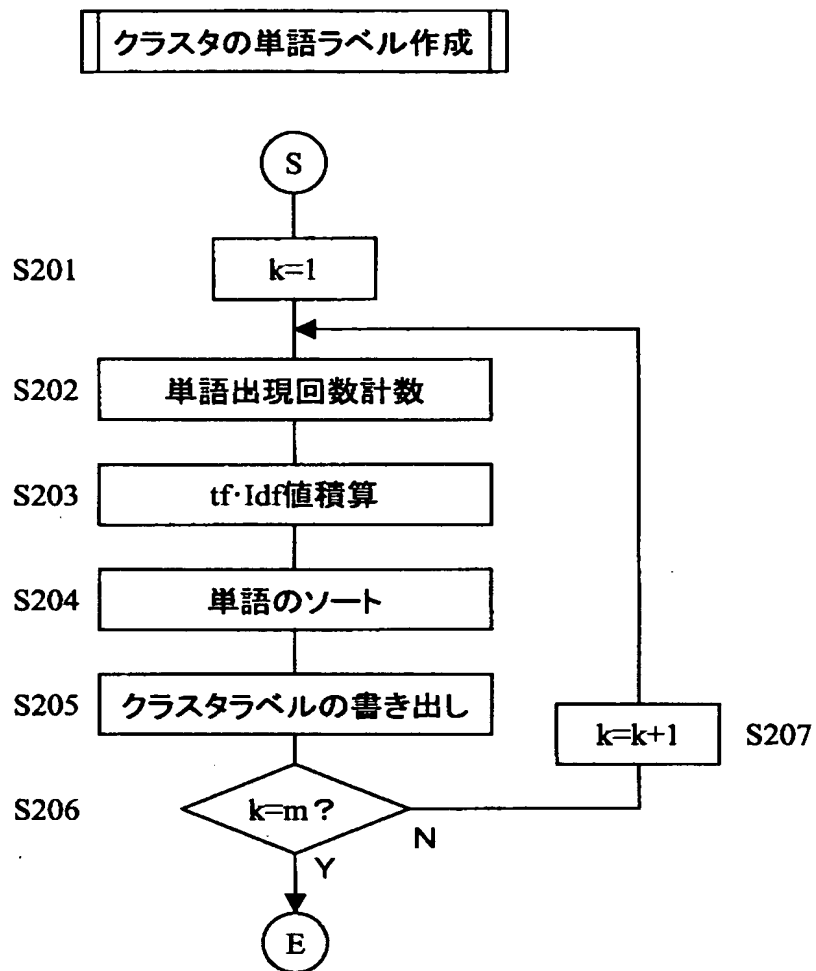
【図 1】



【図 2】



【図 1 0】



【図 1 1】

クラスタID	単語ラベル
1	お菓子, 間食, チーズ
2	体調, エクササイズ, 効果
3	ストレス, 前向き, 状況
4	生理, 食欲, ダイエット
:	:

【図12】

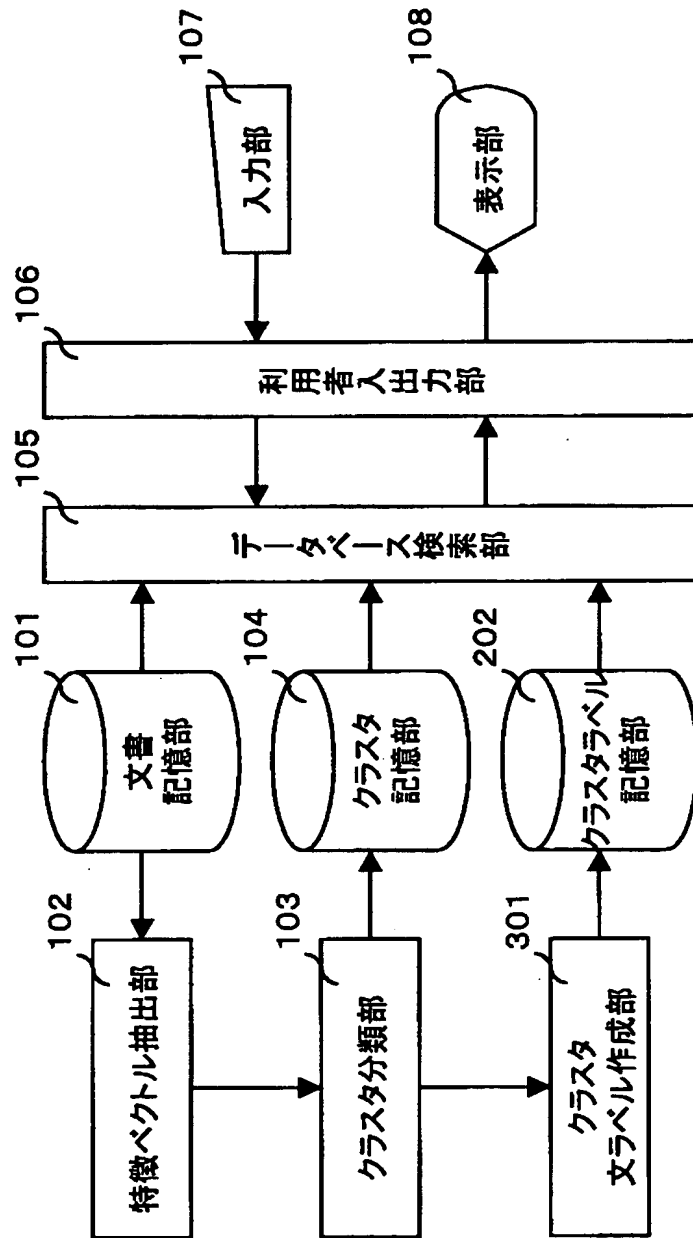
〇〇件の検索結果がありました。

クラスID	単語ラベル	文書ID	文書
1	お菓子 間食 チーズ	1	お菓子が食べなくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。...
		190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、...
		432	間食は200kcal以内で自由に選んでもOKです。...
		644	お菓子が食べなくなったときは、 ・かみごたえがあり、後を引かないもので、...

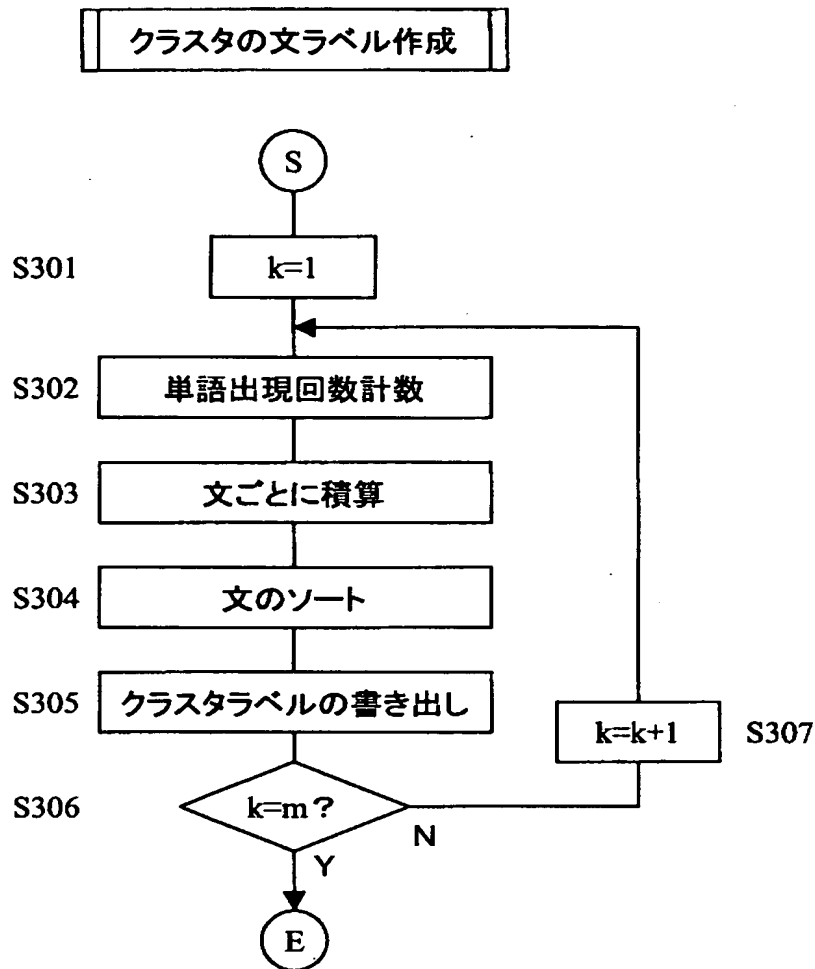
前のクラス

次のクラス

【図 13】



【図 1 4】



【図 1 5】

クラスタID	文ラベル
1	水分の多い物(ゼリー、プリン、ヨーグルト)を...
2	生理中、体調が悪い時には無理にエクササイズを...
3	何がストレスになっているのかを確かめ、...
4	そこで生理前でも自分のダイエットペースを崩さないように...
:	:

【図 16】

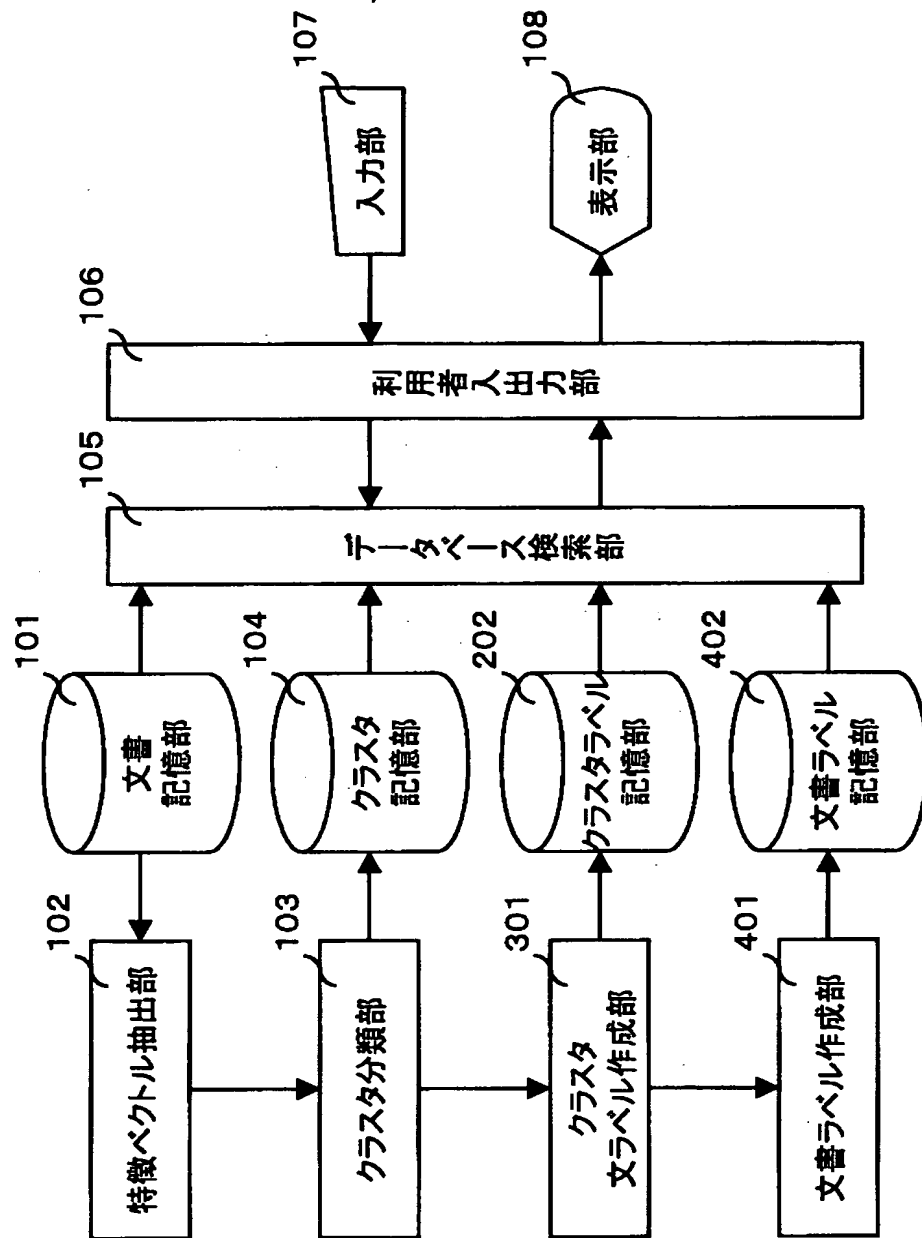
〇〇件の検索結果がありました。

クラスID	文ラベル	文書ID	文書
1	水分の多い 物(ゼリー、 プリン、ヨー グルト)を...	1	お菓子が食べなくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。...
		190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、...
		432	間食は200kcal以内で自由に選んでもOKです。...
		644	お菓子が食べなくなったときは、 ・かみごたえがあり、後を引かないもので、...

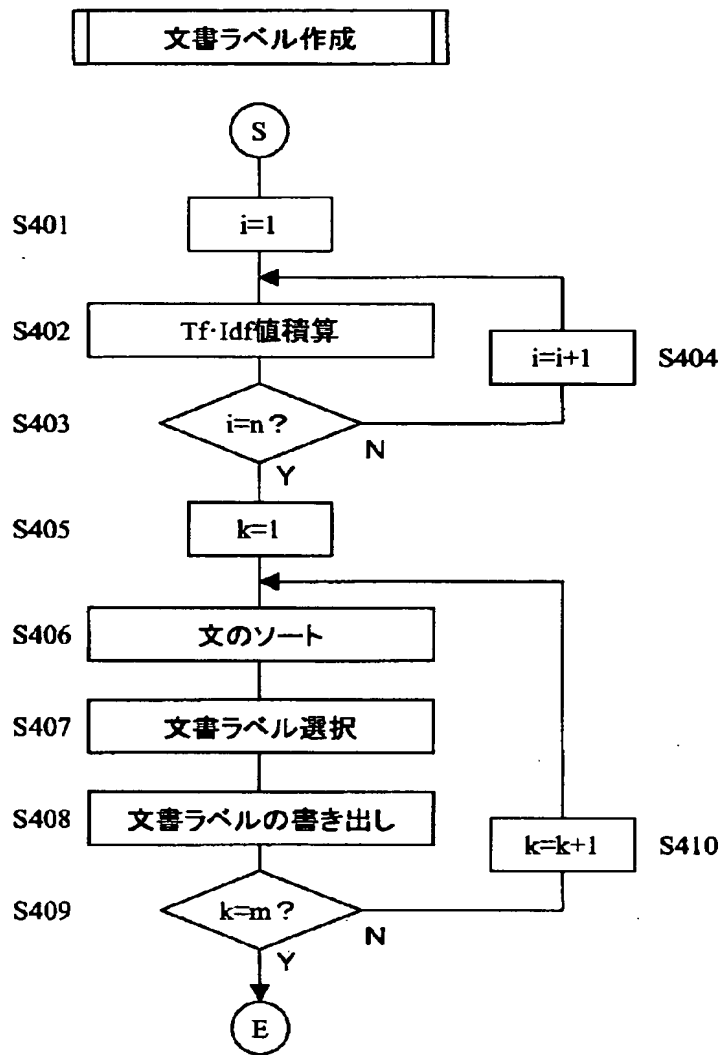
前のクラス

次のクラス

【図 17】



【図 1 8】



【図 19】

クラスID	文書ID	文書ラベル
1	1	かみごたえがあり、後を引かないもので、...
	190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、...
	432	ローカロリーにしたい場合は、低カロリー甘味料の...
	644	どれも、1日のトータルで200kcal以内で、...
:	:	:

【図20】

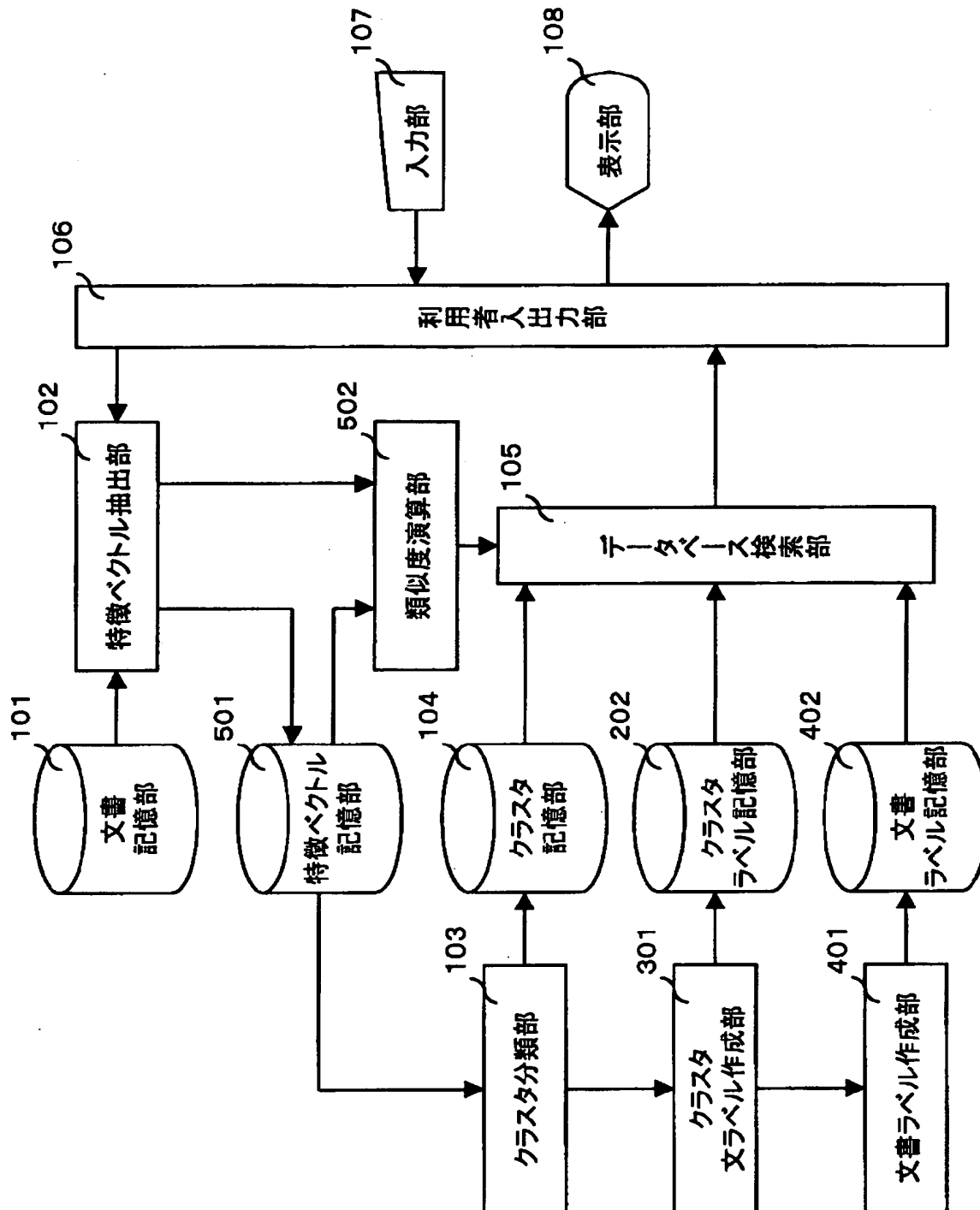
〇〇件の検索結果がありました。

クラスID	文ラベル	文書ID	文書
1	水分の多い 物(ゼリー、 プリン、ヨー グルト)を...	1	お菓子が食べなくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。...
		190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、...
		432	間食は200kcal以内で自由に選んでもOKです。... ローカロリーにしたい場合は、低カロリー甘味料の...
		644	お菓子が食べなくなったときは、 ・かみごたえがあり、後を引かないもので、... どれも、1日のトータルで200kcal以内で、...

前のクラス

次のクラス

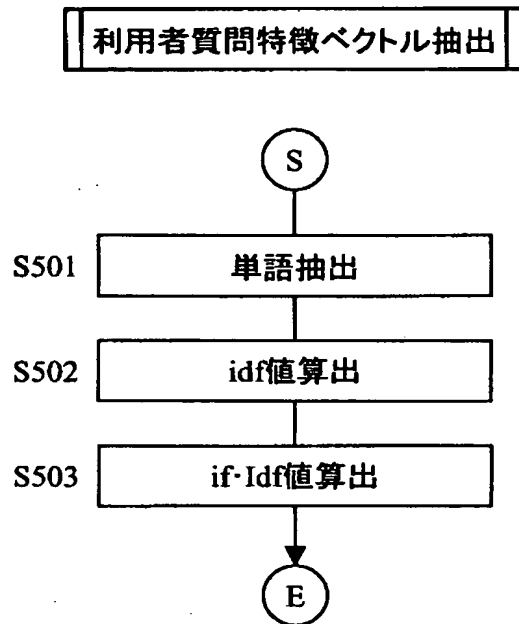
【図 21】



【図 22】

文書ID	質問	回答
1	どうしても、お菓子が食べたくなくなったら、どんなものを食べたいですか。...	お菓子が食べなくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。...
2	朝一番の食事の前の運動が一番効果があると聞いたのですが、夕方の有酸素運動はどうなんですか？...	運動はいつでも、どこでも一人で、が原則です。 人によって生活、体調はみな違います。...
3	今月の生理が1週間以上も遅れてしまってます。...	まったくあなたの考えている通りです。何か悩みがあったり、...
: n個	:	:

【図 2 3】

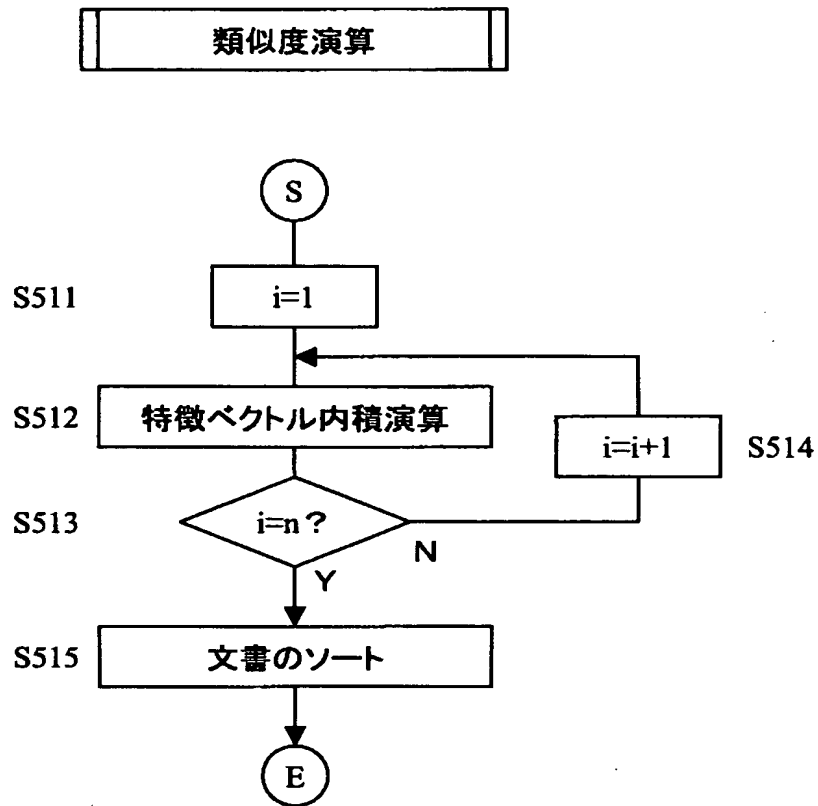


【図 2 4】

特徴ベクトル  $V_Q$  {

単語 $t_i$	重み $w_{Qi}$
お菓子	0.601
間食	0.452
私	0.400
食べる	0.847
朝食	0.556
方法	0.456
⋮	⋮

【図 2 5】



【書類名】 要約書

【要約】

【課題】 従来、情報検索装置や自動質問回答装置では、質問文に類似した文書として、同じような内容の文書が大量に検索されることが多く、利用者は同じような内容の検索結果の中から欲しい情報を見つけ出す必要があったり、あるいは、同じような内容の検索結果が大量に出てきても、自分の欲しい情報がなかったりするという課題があった。

【解決手段】 情報単位の特徴量を抽出し、前記特徴量に基づき前記情報単位のクラスタ分類を行い、前記情報単位から検索条件に合致する情報単位を検索し、検索結果を前記クラスタにまとめて出力する。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000005821]

1. 変更年月日	1990年 8月28日
[変更理由]	新規登録
住 所	大阪府門真市大字門真1006番地
氏 名	松下電器産業株式会社

【図 3】

文書ID	本文
1	お菓子が食べなくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。...
2	運動はいつでも、どこでも一人で、が原則です。 人によって生活、体調はみな違います。...
3	まったくあなたの考えている通りです。何か悩みがあったり、...
: n個	:

文書  $D_i$

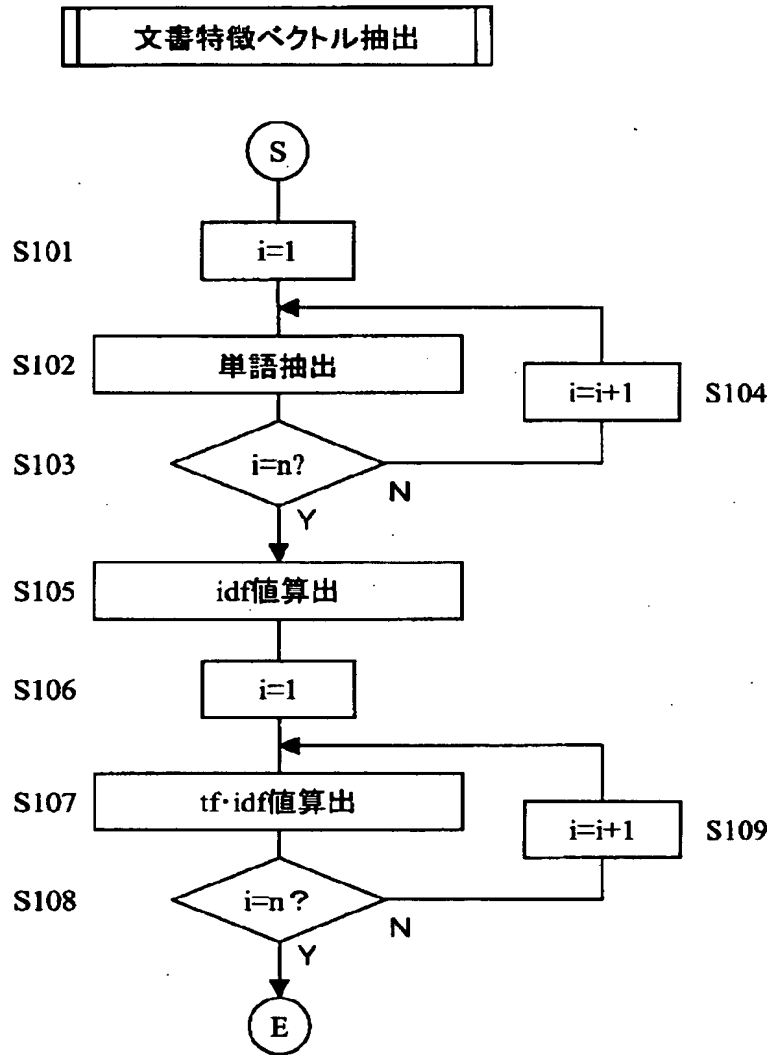
【図 4】

特徴ベクトル  $V_i$

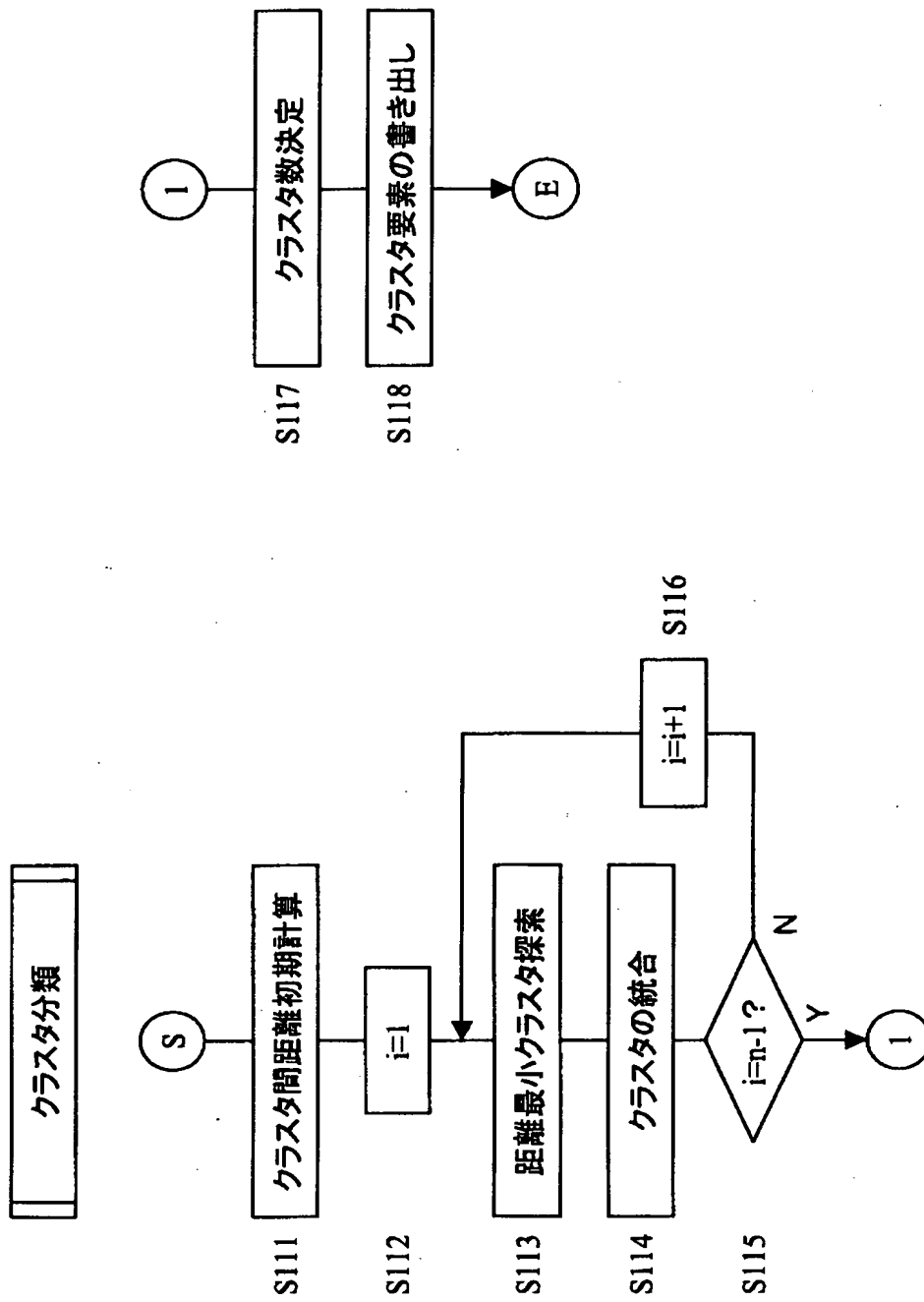
文書ID	単語 $t_{ij}$	重み $w_{ij}$
	お菓子	0.141
	ケーキ	0.193
	スイートポテト	0.223
	ゼリー	0.230
	チーズ	0.173
	とる	0.084
	ノンカロリー	0.161
	プリン	0.223
	飲み物	0.156
	栄養	0.127
	甘いもの	0.174
	食べる	0.231
	食品	0.109
	:	:

n個

【図 5】



【図 6】



【図 7】

クラスタID	文書ID
1	1, 190, 432, 644
2	2, 412, 3, 158
3	3, 158
4	4, 109, 182, 615
:	:

【図8】

〇〇件の検索結果がありました。

クラスタID	文書ID	文書
1	1	お菓子が食べたくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。...
	190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、...
	432	間食は200kcal以内で自由に選んでもOKです。...
	644	お菓子が食べたくなったときは、 ・かみごたえがあり、後を引かないもので、...

前のクラスタ

次のクラスタ

【図9】

